PRIS in TREC 2008 Blog Track

Hui He, Bo Chen, Lei Du, Si Li, Huiji Gao, Weiran Xu, Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
Beijing, P.R. China, 100876
hh1012@gmail.com

Abstract:

This paper describes BUPT (pris) participation in baseline adhoc retrieval task and the opinion retrieval task at Blog Track 2008. The system adopts a two-stage strategy in the opinion retrieval task. In the first stage, the system carries out a basic topic relevance retrieval to get the top 1,000 documents for each query. In the second stage, our system combines several Maximum Entropy based classifiers to implement opinion judging and ranking.

1 Introduction

The Blog track had four tasks in the TREC 2008. We participate in baseline adhoc (blog post) retrieval task and opinion finding (blog post) retrieval task [1]. The baseline adhoc retrieval task involves locating blog posts that contain relevant information about a given topic target. The opinion retrieval task involves locating blog posts that express an opinion about a given target.

The PRIS system submitted by PRIS lab at Beijing University of Posts and Telecommunications adopts a two-stage strategy in the opinion retrieval task. A preprocessing is executed to extract the content from the permalink HTML pages. The basic adhoc retrieval platform is based on the Indri Retrieval Toolkit [2]. In the first stage, the system carries out a basic topic relevance retrieval to get the top 1,000 documents for each query. In the second stage, our system combines several Maximum Entropy based classifiers to implement opinion polarity judging and ranking. These classifiers are trained mainly on the document set extracted from Blog06 corpus according to the relevance judgments for the Blog Track 06&07. After that, such classifiers are applied to the whole Blog06 corpus to generate an opinion (polarity) judgment list for all the documents, combined with the corresponding sentiment strength within interval (0, 1). Finally, in the opinion retrieval task, for each query, the returned top 1,000 documents are re-ranked according to the score consisting of the topic relevance and the opinion sentiment strength.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
PRIS in TREC 2008 Blog Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Beijing University of Posts and Telecommunications,School of Information and Communication Engineering,,Beijing, P.R. China, 100876,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**5**	

The remainder of this paper is organized as follow. In section 2, a briefly system overview is presented. Section 3 describes the preprocessing part. Section 4 introduces the topic retrieval part. Section 5 presents the opinion finding part. Evaluation results are shown in section 6.

2 System Overview

The PRIS system consists of preprocessing part, topic retrieval part and opinion finding part. Architecture of PRIS is shown in Figure 1.

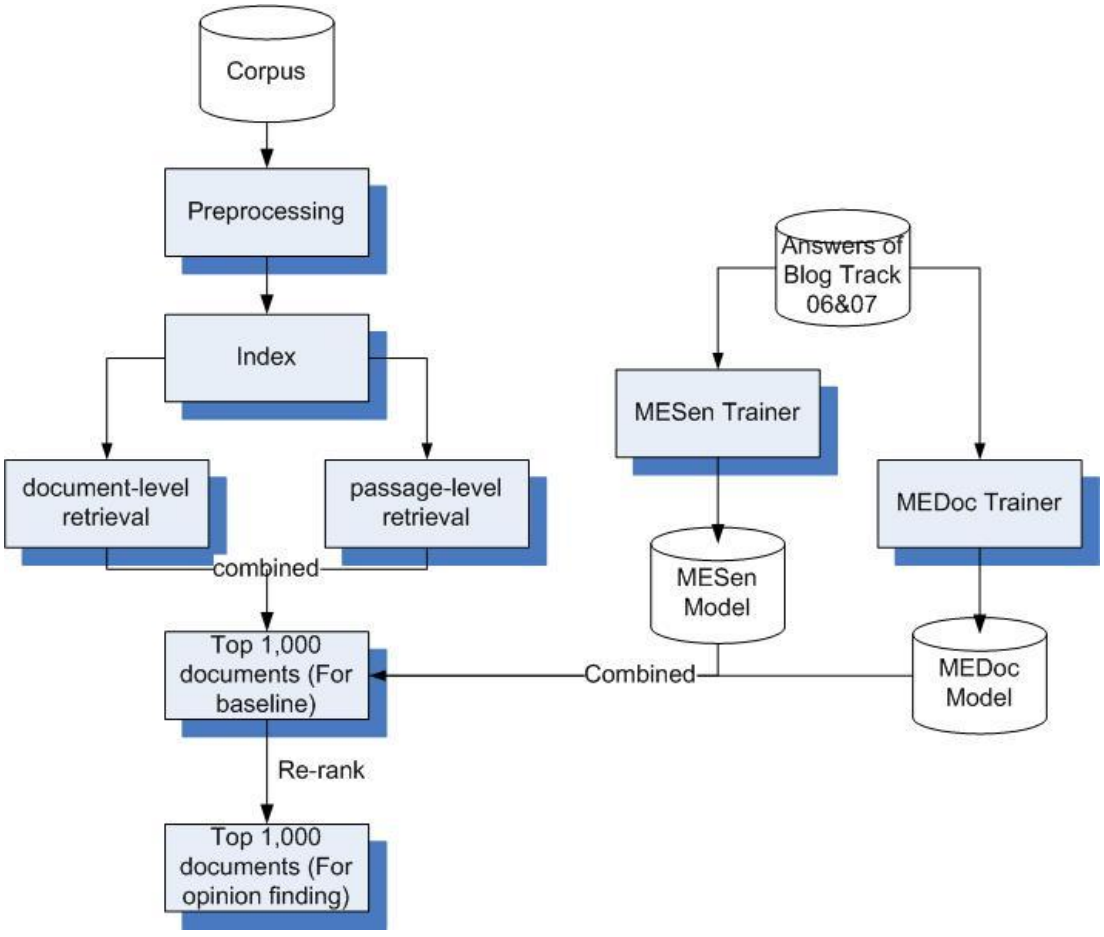

Figure 1. System Architecture

The preprocessing part is designed to extract content of the permalink HTML pages, and some rules are set to process abbreviations.

The topic retrieval part based on the Indri Retrieval Toolkit tires both the document-level retrieval and the passage-level retrieval, and then, such two returned lists are combined to a final query result list. Top 1,000 documents for each query are selected to be re-ranked.

The opinion finding part contains two Maximum Entropy based classifiers. One is sentence-level trainer, which generates ME sentence-level models to rank and judge sentence polarity. Another is document-level trainer, which generates ME document-level modes to rank and judge document sentiment orientation. Document-level trainer makes use of sentence labels judged by ME sentence-level models.

Finally, top 1,000 documents are re-ranked according to the score consisting of the topic relevance and the opinion sentiment strength.

3 Preprocessing

The corpus contains permalinks HTML pages, feed files and blog homepages.

We only use the permalink HTML pages for retrieval. These HTML pages are parsed and texts are reserved. The hyper-links, scripts, style information in the web pages and all html tags are discarded. After the preprocessing, the cleaned permalink HTML pages amounted to about 15G.

In addition, we apply some rules to cleaned pages to abbreviations such as:

“I’m”, “They’re” to be processed to “I am”, “They are”

We don’t apply word stem on the corpus. Such heuristic rules in content preprocessing are useful for topic retrieval task and opinion finding retrieval task.

4 Topic Retrieval

Based on the preprocessed corpus, index is build using Indri. For most of the 150 queries, only items in the "title" field are used as the query inputs with some automatic/manual reformulations, such as Google-based query expanding, term-weighting, structural query and so on. Some expanded items are from “desc” field and “narr” field. For example, for topic “jstor”, the query is as follow after query expansion and structuration.

```
<query>
<num> Number: 901 </num>
<text>#weight(1.0 jstor 0.001 journals 0.001 archive 0.1 #uw40(difficult jstor) 0.1
  #uw40(ease jstor)0.2#not(opinion) 0.2#not(reference))</text>
</query>
```

Also, we have tired both the document-level retrieval and the passage-level retrieval, and then, such two returned lists are combined to a final query result list. We get the top 1,000 documents for each query.

5 Opinion Finding Retrieval

The second stage is opinion finding retrieval. We employ a hierarchical sentiment analysis model based on Maximum Entropy classifier [3]. The hierarchical sentiment analysis model is shown in Figure 2.

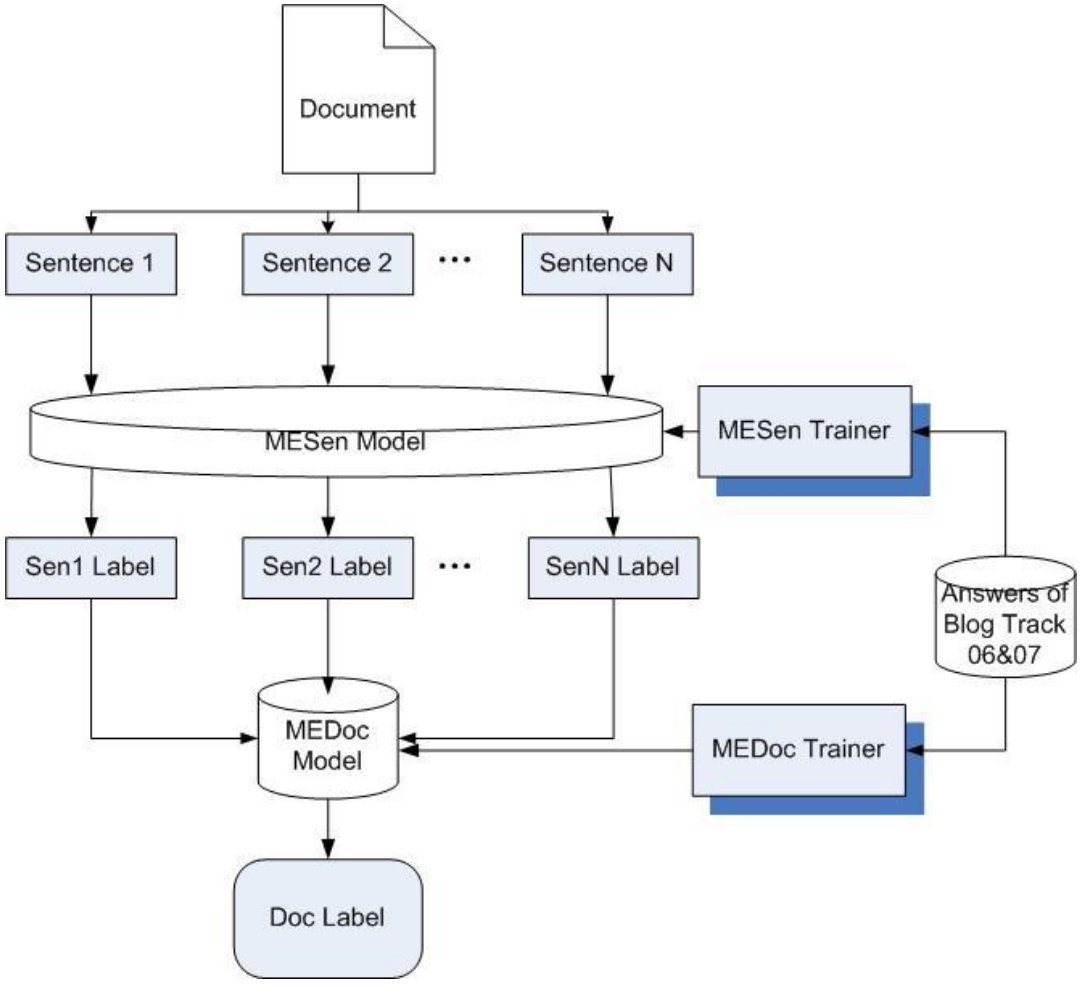

Figure 2. Hierarchical Sentiment Analysis Model

Answers of Blog Track 06&07 are used as training data. MESen trainer generates sentence-level models "MESen models". MEDoc trainer makes use of sentence sentiment sequence information and generates document-level models "MEDoc model". For example, a document contains several sentences. Each sentence is classified by MESen model and gets a label for its sentiment orientation. After that, this document has a sequence of sentence sentiment labels. MEDoc models judge and label such sequence.

Such hierarchical sentiment analysis model is applied to the whole Blog06 corpus to generate an opinion (polarity) judgment list for all the documents, combined with the corresponding sentiment strength within interval (0, 1). For each query, the returned top

1,000 documents are re-ranked according to the score consisting of the topic relevance and the opinion sentiment strength.

6 Submission and Evaluation Results

We submitted 4 runs, as follows:

- prisba: A baseline run only using query of title field and automatically retrieved by Indri.
- prisbm: Run with query expansion based on Google query expanding and manually term-weighting.
- prisoa1: the prisba run with opinion finding retrieval.
- prisom1: prisbm run with opinion finding retrieval.

The evaluation results of the 4 submitted runs are listed in table 1 and table 2.

Table 1. Topic Relevance Results

Run	MAP	R-prec	b-Pref	P@10
prisba	0.4245	0.4733	0.5122	0.7113
prisbm	**0.4424**	**0.4868**	**0.5411**	0.7793
prisoa1	0.4243	0.4727	0.5119	0.7100
prisom1	0.4418	0.4863	0.5408	**0.7807**

Table 2. Opinion Finding Results

Run	MAP	R-prec	b-Pref	P@10
prisba	0.3103	0.3644	0.3573	0.5167
prisbm	0.3147	0.3709	0.3722	0.5307
prisoa1	0.3105	0.3656	0.3575	0.5193
prisom1	**0.3149**	**0.3721**	**0.3723**	**0.5347**

References

[1] Craig Macdonald, Iadh Ounis, and Ian Soboroff. “Overview of the TREC-2007 Blog Track”, In TREC 2007, 2007.

[2] Indri search engine package: http://www.lemurproject.org/indri/.

[3] Bo Chen, Hui He, Jun Guo. “Constructing Maximum Entropy Language Models for Movie Review Subjectivity Analysis”. Journal of Computer Science and Technology (JCST), 23(2): p231-239, 2008.